

Sklop 3: Klasifikacija Naive Bayes v orodju RapidMiner

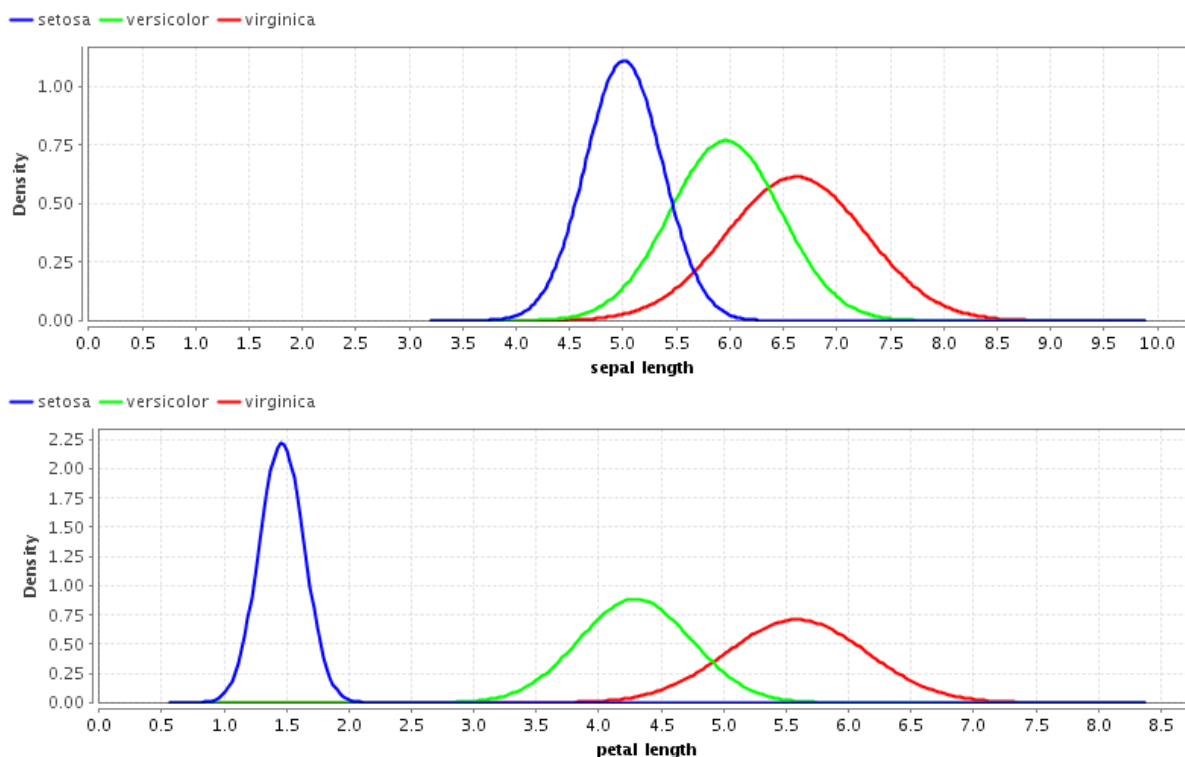
Uvod

Klasifikacija vzorcev je metoda strojnega učenja, pri kateri neznani vzorec razvrščamo v razred, glede na njegovo podobnost vzorcem z znano razvrstitvijo. V postopku potrebujemo t.i. učno množico, ki vsebuje vzorce, katerim je pripisana razvrstitev. Z algoritmi strojnega učenja iz učne množice zgradimo model, ki ga kasneje uporabimo za razvrščanje neznanega vzorca.

Obstaja mnogo metod za klasifikacijo vzorcev, uporabili bomo klasifikator imenovan Naivni Bayesov klasifikator. Klasifikator je preprost verjetnostni klasifikator, ki za vsako značilko za posamezen razred izračuna verjetnostno porazdelitev le te. Neznani vzorec razvrsti tako, da najprej izračuna verjetnosti, da posamezen vzorec pripada razredu glede na porazdelitveno funkcijo posameznih značilk in te verjetnosti združi. Končno razvrstitev poda po kombinaciji posameznih verjetnosti – odloči se za tisto, pri kateri je kombinirana verjetnost največja.

Pri uporabi tovrstnega klasifikatorja se je potrebno zavedati, da je predpostavljena neodvisnost posameznih značilk.

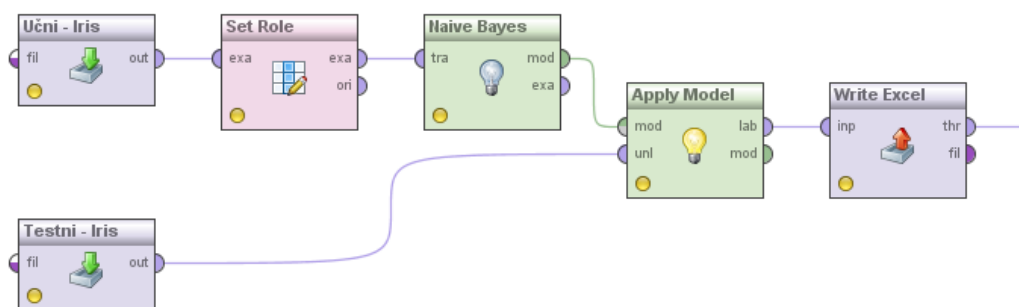
Na sliki sta prikazani dve od štirih verjetnostnih porazdelitev značilk podatkovnega nabora Iris za posamezne vrste, izračunane s programom RapidMiner.



Slika: Verjetnostni porazdelitvi dveh od štirih značilk za učno množico Iris.

Naloga

- 1) Z uporabo Operatorja **Read Excel** naložite učne podatke z zavihka Iris-učni
- 2) Z uporabo Operatorja **Read Excel** naložite testne podatke z zavihka Iris-testni
- 3) Oglejte si podatke. Narišite raztresne grafe podatkov Iris-učni.
 - Scatter plot (x:petal length, y: petal width, Color: species)
 - Scatter matrix (Plots:species)
 - Scatter 3D Color (x: sepal length, y: petal length, z: petal width, Color: species)
- 4) V učnih podatkih z uporabo operatorja **Set Role** definirajte vlogo podatkov: za značilko »species« definirajte vlogo label
- 5) Z uporabo operatorja **Naive Bayes** zgradite model na učnih podatkih. Oglejte si model.
- 6) Z uporabo operatorja **Apply Model** uporabite naučeni model na testnih podatkih. Oglejte si izračunane napovedi modela za testne podatke. Rezultate primerjajte s podatki v Excelovi tabeli Iris-kontrola.
- 7) Rezultate napovedi shranite v datoteko z uporabo operatorja **Write Excel**.



Slika: Proces nalaganja učne in testne množice podatkov, predobdelave, gradnje modela in uporabe le tega na testni množici v orodju RapidMiner.

Vir

-Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining. Practical Machine Learning. Tools and Techniques. Third Edition

- <http://rapid-i.com/>

- <http://www.statsoft.com/textbook/naive-bayes-classifier>

- http://en.wikipedia.org/wiki/Iris_flower_data_set

Vaja pripravljena v okviru KCUI delavnice na UL-FE